AD-A105 906 PRINCETON UNIV NJ DEPT OF STATISTICS F/G 12/1
SOME ADVANCE THOUGHTS ON THE DATA ANALYSIS INVOLVED IN CONFIGUR--ETC(U)
MAR 81 J W TUKEY DAAG29-79-C-0205
UNCLASSIFIED TR-189-SER-2 ARO-16669.8-M NL

1 OF 1
AD A
105906

END
DATE
FILMED
11-81
DTIC

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

Level II

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
19 16669.8-M 18 ARO

2. GOVT ACCESSION NO.
AD-A105

3. RECIPIENT'S CATALOG NUMBER
906

4. TITLE (and Subtitle)
6 Some Advance Thoughts on the Data Analysis Involved in Configural Polysampling Directed Toward High Performance Estimates .

5. TYPE OF REPORT & PERIOD COVERED
9 Technical rept.

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
10 John W. Tukey

12

8. CONTRACT OR GRANT NUMBER(s)
15 DAAG29-79-C-0205

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Princeton University
Princeton, NJ 08544

12 24

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

12. REPORT DATE
11 16 Mar 81

13. NUMBER OF PAGES
17

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
14 TR-189-SER-2

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Using configural polysampling to identify high-performance estimates proceeds in two phases: (1) estimating the best compromise to make at each configuration polysampled, and (2) choosing a generally-defined estimate whose performance comes close to that of this estimated optimum. In the first step, shadow pricing of the criteria for the differenc situations involved is helpful. In the second step, whether one uses a regression or selection-by-sectors approach, or combine both, the same shadow prices are useful. This report sketches some of the approaches to the two steps that seem natural.

UNCLASSIFIED 406873 LB

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Some advance thoughts on the data analysis involved in configural polysampling directed toward high performance estimates*

by

John W. Tukey
Princeton University
Princeton, New Jersey 08544
and
Bell Laboratories
Murray Hill, New Jersey 07974

Technical Report No. 189, Series 2
Department of Statistics
Princeton University
March 1981

*Prepared in part in connection with research at Princeton University sponsored by the Army Research Office (Durham) and in part in connection with research at Bell Telephone Laboratories.

ABSTRACT

Using configural polysampling** to identify high-performance estimates proceeds in two phases:

- estimating the best compromise to make at each configuration polysampled,
- choosing a generally-defined estimate whose performance comes close to that of this estimated optimum.

In the first step, shadow pricing of the criteria for the different situations involved is helpful.

If the criteria are the relative excess variances for the situations in question, minimizing their maximum corresponds to maximizing polyefficiency. In this case, as in most others, finding the shadow prices is likely to require iteration.

In the second step, whether we use a regression or selection-by-sectors approach, or combine both, the same shadow prices are useful.

This report sketches some of the approaches to the two steps that seem natural.

**See Technical Report No. 185 by Pregibon and Tukey for the ideas of configural polysampling. See Technical Report No. 191 by Bell and Pregibon for an implementation.

1. Introduction to A and B

One of the purposes of configural polysampling is to identify -- on the configurations sampled, at least -- estimates of high polyperformance, and, as well, to assess the extreme possibilities of specific kinds of polyperformance. Often these polyperformances are functions of the various sampling variances, functions of such a character that the family of Sayeb estimates -- of estimates minimizing

$$\Sigma \left[(\text{shadow price for } (Q)) \cdot \frac{(\text{variance at } Q) - (\text{constant})}{(\text{constant}^*)} \right]$$

for some set of shadow prices -- includes, by necessity, the estimate with highest polyperformance.

Below, we first discuss, in some generality, the evaluation of Sayeb estimates in terms of shadow prices and then, second, a plausible approach to iterative selection of the shadow prices in the special case where polyefficiency is the polyperformance to be maximized. In this case, of course, we can equally well treat the largest relative excess variance

$$\max_{Q} \frac{(\text{variance at } Q) - (\text{minimum variance at } Q)}{\text{minimum variance at } Q}$$

as a polyperformance to be minimized.

We shall work with configurations specified as

Prepared in part in connection with research at Princeton University, sponsored by the U. S. Army Research Office (Durham), and in part in connection with research at Bell Telephone Laboratories.

March 16, 1881

$c_1, c_2, \ldots, c_n$, where $c_a=0$, $c_b=1$ and

$$y_i = r + c_i s$$

where $y_1 \leq y_2 \leq \cdots \leq y_n$ are the order statistics of a sample, and a and b are likely to be chosen close to n/4 and 3n/4. (Note that r and s are coordinates locating a particular sample in the configuration; they do <u>not</u> have their most customary statistical meanings.)

A

The Sayeb minimizers

2. The case of two situations.

We start with only two situations, A and Z, and any one configuration. If t locates the estimate relative to the configuration, as when T=r+ts, the conditional mean square error of the estimate, for this configuration, is a quadratic function of t. We lose nothing if we modify the scale of t so that t = 0 is the optimum value of t in situation A and t = 1 is optimum for situation Z.

If T_{fA} and T_{fZ} were the original optimum values, any value of t would correspond (where r and s locate a sample in the configuration f) to

$$\overset{f}{\underset{r,s}{\mathrm{Var}}}(r+(T_{fA}+t(T_{fZ}-T_{fA}))s)$$

$$= (\text{minimized A-variance})_f + t^2(T_{fZ}-T_{fA})^2\overline{s^2_{fA}}$$

and thus to a contribution to the weighted sum of excess variances at situation A equal to

$$t^2(W_{Af}\overline{s^2_{fA}}(T_{fZ}-T_{fA})^2)$$

where W_{Af} is the weight on configuration f appropriate for situation A. (In dealing with operators like "var" which may involve a restricted range, we specify the range <u>above</u> the operator, reserving the space below the operator for binary variables. Thus the "f" above the "var" above indicates a conditional variance over the configuration f.) (The "weights" used are discussed in Technical Report #185.)

If we are concerned with polyefficiency, we are concerned with the excess-variance as a fraction of the minimum variance (minimum when A alone is considered) which we call the "minimum A-variance". Thus our concern, at f under A, about t differing from zero is the shadow price times

$$t^2\left|\frac{W_{Af}\overline{s^2_{fA}}(T_{fZ}-T_{fA})^2}{\text{minimum A-variance}}\right| = a_{Af}t^2$$

while that at f under Z is

$$(1-t)^2 \left| \frac{W_{Zf}\overline{s^2_{fZ}}(T_{fZ}-T_{fA})^2}{\text{minimum Z-variance}} \right| = a_{Zf}(1-t)^2$$

If the shadow prices are α and ξ, we then wish to minimize

$$\alpha \sum_f a_{Af}t_f^2 + \xi \sum_f a_{Zf}(1-t_f)^2$$

Differentiation w.r.t. t_f leads to

$$\alpha a_{Af}t_f - \xi a_{Zf}(1-t_f) = 0$$

and thence to

$$t_f = \frac{\xi a_{Zf}}{\alpha a_{Af} + \xi a_{Zf}}$$

and hence to a total price that is the sum over f of

$$\frac{(\alpha a_{Af})(\xi a_{Zf})}{\alpha a_{Af} + \xi a_{Zf}} .$$

More interesting, usually, will be the individual excess variances

$$\Sigma\ a_{AF}t_f^2 = \Sigma\ a_{Af}\ \frac{\xi^2 a_{Zf}^2}{(\alpha a_{Af} + \xi a_{Zf})^2}$$

$$= \xi^2 \Sigma\ a_{Zf}\ \frac{a_{Af}a_{Zf}}{(\alpha a_{Af} + \xi a_{Zf})^2}$$

and

$$\Sigma\ a_{Zf}(1-t_f)^2 = \Sigma\ a_{Zf}\ \frac{\alpha^2 a_{Af}^2}{(\alpha a_{Af} + \xi a_{Zf})^2}$$

$$= \alpha^2 \Sigma\ a_{Af} \cdot \frac{a_{Af}a_{Zf}}{(\alpha a_{Af} + \xi a_{Zf})^2}$$

If we introduce

$$r_f = \frac{\alpha a_{Af}}{\xi a_{Zf}}$$

and

$$\phi(r) = \frac{4}{r + 2(1/r)} = \mathrm{sech}^2(\ln r)$$

which = 1 if $r = 1$, about 1/3 for $r = 0.1$, or 10, and about 1/25 for $r = .01$ or 100, these the excess variances become

$$\Sigma\ a_{Af}t_f^2 = (1/4)(\xi/\alpha)\ \Sigma\ a_{Zf}\phi(r_f)$$

$$\Sigma\ a_{Zf}(1-t_f)^2 = (1/4)(\alpha/\xi)\ \Sigma\ a_{Af}\phi(r_f)$$

which will be less than, but perhaps of the order of

$$\frac{\xi}{4\alpha}\ \Sigma\ a_{Zf}$$

and

$$\frac{\alpha}{4\xi}\ \Sigma\ a_{Af}$$

What if there are more than two situations?

3. Three situations.

Suppose now that we consider an additional situation, B, one where the relative excess variance is estimated by

$$\sum_f a_{Bf}(t-b)^2$$

where t = b (not necessarily between 0 and 1) is the optimizing value of t for situation B alone. With shadow prices α,β and ξ we are to minimize

$$\Sigma\ (\alpha a_{Af}t^2 + \beta a_{Bf}(t-b)^2 + \xi a_{Zf}(t-1)^2)$$

We can simplify this by using

$$(1-c)t^2 + c(1-t)^2 - c(1-c) \equiv (t-c)^2$$

an identity easily checked. Changing notation, we have

$$\beta a_{Bf}(t-b)^2 \equiv \beta a_{Bf}(1-b)t^2 + \beta a_{Bf}b(t-1)^2 - \beta a_{Bf}b(1-b)$$

and what we are to minimize becomes

$$\Sigma \left| (\alpha a_{Af} - \beta a_{Bf}(1-b))t^2 + (\xi a_{Zf} + \beta a_{Bf}b)(t-1)^2 - \Sigma\, \beta a_{Bf}b(1-b) \right|$$

where the final term is independent of t.

By analogy with the last section, we see that

$$t_f = \frac{\xi a_{Af} + \beta b a_{Bf}}{\alpha a_{Af} + \beta a_{Bf} + \xi a_{Zf}}$$

and we may as well calculate the three excess variances as

$$E_A = \Sigma\, a_{Af} t_f^2$$

$$E_B = \Sigma\, a_{Bf}(t_f - b)^2$$

$$E_Z = \Sigma\, a_{Zf}(t_f - 1)^2$$

respectively.

4. Four or more situations.

The formulas for four situations follow in an entirely similar way. From them the general relations lead to the Sayeb minimizations $(E_A, E_B, E_C, \ldots, E_Z)$ given by

$$E_A = \Sigma\, a_{Af} t_f^2$$

$$E_B = \Sigma\, a_{Bf} (t_f - b)^2$$

$$E_C = \Sigma\, a_{Cf} (t_f - c)^2$$

$$\ldots \quad \ldots$$

$$E_Z = \Sigma\, a_{Zf} (t_f - 1)^2$$

where

$$t_f = \frac{\beta b a_{Bf} + \gamma c a_{Cf} + \ldots + \xi a_{Zf}}{\alpha a_{Af} + \beta a_{Bf} + \gamma a_{Cf} + \ldots + \xi a_{Zf}}$$

B

A suggested iteration

5. Worst-case minimization.

We now suppose that our desire is to minimize

$$\max\{E_A, E_B, E_C, \ldots, E_Z\}$$

and that we have tried $\alpha(0), \beta(0), \ldots, \xi(0)$ with results $E_B(0)$, $E_B(0)$,..., $E_Z(0)$, whose maximum is $E_{max}(0)$.

If $E_A(0)$ is less than $E_{max}(0)$, we are doing wastefully well at A, so that we want to reduce the corresponding relative shadow price, α, and let E_A increase in the hope that E_{max} decreases. Similarly if $E_B(0)$ is less than $E_{max}(0)$, we would like β to sink relatively. And so on. Let us propose one algorithm that moves things in the right direction.

For some exponent p -- we try p = 1 (or, perhaps, p = 0) for awhile, and then learn to choose p more reasonably -- and some convenient k, we can take

$$\alpha(1) = k\left|\frac{E_A(0)}{E_{max}(0)}\right|^p \alpha(0)$$

$$\beta(1) = k\left|\frac{E_B(0)}{E_{max}(0)}\right|^p \beta(0)$$

$$\cdots \quad \cdots$$

$$\xi(1) = k\left|\frac{E_Z(0)}{E_{max}(0)}\right|^p \xi(0)$$

(It might be nice to choose k so that $\alpha(1)+\beta(1)+\ldots+\xi(1)=1000$ -- or some other handy number.)

Looking at a few successive values of $\{(\alpha(i),\beta(i),\ldots,\xi(i))\}$ might now suggest, for instance, trying certain of these shadow prices as zero. If, while using them as zero, the corresponding E ever becomes E_{max}, we should make them finite (>0) again.

This may not be the most rapidly convergent algorithm, but we can hope for relatively good performance, once we tune p.

C

External description of estimates

6. Introduction to C.

We have discussed how, in simple instances we can estimate internal descriptions of "best" estimates, finding out -- for a sample of configurations -- good estimates of what the value of the "best" estimate is for that configuration. (The failure of our "estimates" to be perfect comes from our slight imprecision in finding the right shadow prices.)

We now want to estimate an external description of the "best" estimate. This means that we need to find a function, defined for all configurations that comes close to our estimate values for each of our sample configurations. Several approaches are possible including:

1) Selecting a set of explicit estimates, evaluating each at each selected configuration, and doing a linear regression of our "best" values on the estimate values -- the only question is what weights to use in the regression,

2) dividing the space of configurations up into parts, which we will call sectors, evaluating the performance of each of the explicit estimates for each sector (using excess variances and the optimizing shadow prices), and selecting the explicit estimate that performs best in each sector,

3) continuing (2) by trying to interpolate sensibly between the several explicit estimates thus selected, one for each sector,

4) combining (1) and (2), which leads, after help from experts, to a solvable linear programming problem, and to different lincoms in different sectors

5) taking a relatively good explicit estimate, identifying (in terms of excess variance at optimizing shadow prices) its unnecessary loss at each configuration, isolating the large-loss configurations, and studying the change in estimate required at each of them as a basis for inventing possibly improved estimates.

We will have to learn by experience which of these seem most effective. A little consideration will clear up certain aspects of what we might do.

7. The regression approach.

Suppose, then, that we have, for a sample of configurations f, and two situations, A and Z:

- The overall shadow prices α and ξ that minimize the maximum relative excess variance.
- for each f, the constants a_{Af}, a_{Zf} and t_f (the latter on the modified f-scale where t=0 and t=1 are the separate optima and do not correspond to y_a and y_b)
- for each of m explicit estimates the values t_{fi} for the i-th estimate of t.

Take first m = 2. We want to consider

$$\theta(\text{estimate } 1) + (1-\theta)\ (\text{estimate } 2)$$

whose t-value at f is

$$\theta t_{f1} + (1-\theta)t_{f2}$$

The shadow-priced cost of this estimate is

$$\alpha\ \Sigma\ a_{Af}(\theta t_{f1}+(1-\theta)t_{f2})^2 + \xi\ \Sigma\ a_{Zf}(1-\theta t_{f1}-(1-\theta)t_{f2})^2$$

which equals

$$\alpha\theta^2\ \Sigma\ a_{Af}t_{f1}^2 + 2\alpha\theta(1-\theta)\ \Sigma\ a_{Af}t_{f1} + \alpha(1-\theta)^2\ \Sigma\ a_{Af}t_{f2}^2$$

$$+\ \xi\theta^2\ \Sigma\ a_{Zf}(1-t_{f2})^2 + 2\xi\theta(1-\theta)\ \Sigma z_{Zf}(1-t_{f1})(1-t_{f2})$$

$$+ \xi(1-\theta)^2 \Sigma a_{Zf}(1-t_{f2})^2$$

$$= \alpha\theta^2[A11] + 2\alpha\theta(1-\theta)[A12] + \alpha(1-\theta)^2[A22]$$

$$+ \xi\theta^2[Z11] + 2\xi\theta(1-\theta)[Z12] + \xi(1-\theta)^2[Z22]$$

where the six[ ]'s designate the six sums written down above. If we write

$$[11] = \alpha[A11] + \xi[Z11]$$

$$[12] = \alpha[A12] + \xi[Z12]$$

$$[22] = \alpha[A22] + \xi[Z22]$$

where the first and last are the prices of the individual estimates, the price of the linear combination is

$$\theta^2[11] + 2\theta(1-\theta)[12] + (1-\theta)^2[22]$$

whose θ derivative is

$$2\theta[11] + 2(1-2\theta)[12]-2(1-\theta)[22]$$

which vanishes when

$$\theta_0 = \frac{[22] - [12]}{[11] - 2[12] + [22]}$$

so that

$$1 - \theta_0 = \frac{[11] - [12]}{[11] - 2[12] + [22]}$$

and the price of the optimum linear combination reduces to

$$([11][22])-[12])^2([11]+[22]) - 2[12][11][22] + 2[12]^3$$

divided by the square of

$$[11] - 2[12] + [22]$$

Extending this numerically to larger m offers little difficulty, since what we are doing is equivalent to *seeking a minimum "variance" when*

$$[Zii] = \Sigma\, a_{Zf}(1-t_{fi})^2$$

$$[Zij] = \Sigma\, a_{Zf}(1-t_{fi})(1-t_{fi})$$

The extension to two estimates and more than two situations proceeds in a similar way. There will be more shadow prices and more sums, but, after pricing out, the same number of "variances" and "covariances".

8. Sector techniques.

If we were working with situations

A = Gaussian

Z = Slash

we would probably want to define sectors in terms of some measures of apparent Gaussiantity-vs.-stretchedness, such as

K_1 = javaj ln s^2

$K_2 = \ln s_{bi}^2 - \ln s^2$

$K_3 = c_n - c_1$

$K_4 = c_{n-1} - c_2$

where "javaj" stands for "jackknife-estimate of variance of the jackknifed values of", s_{bi}^2 is the robust estimate of variability described, inter alia, in Chapter 11 of Mosteller and Tukey, and $\{c_1, c_2, \ldots, c_n\}$, are order statistics of the configuration in question.

We might want to use such measures either alone or in combination with some measures of skewness, such as

L_1 = (abmean of c_i's) - (mean of c_i's)

L_2 = (abmean of c_i's) - (median of c_i's)

$L_3 = c_n + c_1$

$$L_4 = c_{n-1} + c_2$$

where the abmean is the mean of the --- central order statistics (out of n).

Having chosen sectors, we now need to estimate the performance of each external estimate at the configurations in our polysample that fall in each sector. When we need to look, for each estimate and sector:

- for each configuration, at the contributions to relative excess variance, say, $a_{Af}t_{fi}^2$ and $a_{Zf}(1-t_{fi})^2$) (recall that the a's incorporate the relevant weights),

- for each sector, at the priced-out contributions

  $$\alpha \Sigma^* a_{Af}t_{fi}^2 + \xi \Sigma^* a_{Zf}(1-t_{fi}^2)^2$$

  where the sums are over all configurations in the <u>sector</u> from our <u>polysample</u>.

If we are selecting estimates, we need now only select for each sector the estimate whose sector price is least. (We may be able to make good use of a listing of excesses of these estimated sector prices over the similar estimate for our optimum as defined by the t_{fi}.)

To go a useful step further, we might list the estimates in order of increasing sector prices for that sector, and then try, in each sector, all 50-50 combinations of the best 10 (or best 6) estimates so far tried for that sector. Sufficient iteration here should lead to close to the optimum linear combination for each sector.

A regression calculation in each sector would also be feasible.

Everything is easy, so long as we are prepared to stick to the previously found shadow prices. We will only get into linear programming situations etc. if we insist on using more correct shadow prices -- more correct because we are no longer accommodating as wide a variety of possible choices, thus shrinking the attainable choice set and tilting the tangent hyperplane at the new optimum This corresponds to insisting on controlling, for our external estimates, the actual

maximum relative excess variances

for the more restricted situation, instead of controlling that particular

lincom relative excess variances

whose unrestricted optimum over our sampled configurations enforces an optimum for the maximum. Those who wish to modify may do so.

9. Use of judgment, intuition, insight and inventiveness

Approaches (3) and (5) above call for human input. We will have to learn how to do this by trying.

DAT

ILM